Magda's notes about information bottleneck in learning

Last update: December 27, 2019

Informal notes for my future self who is likely to forget. I explain the papers the way I understand them, using terminology and logic natural to me. This means I may deviate from the original paper structure, notation, etc. At places, my interpretation my be incorrect due to lack of understanding. I will strive for this not to happen too often but I'm certainly not infallible.

This is a working document, not polished, with possible typos, editing errors, etc.

Contents

1 Tishby's information bottleneck method

Paper: Naftali Tishby, Fernando C. Pereira, et al. "The Information Bottleneck Method". In: *arXiv:physics/0004057* (2000)

Notes taken: 25/12/2019

1.1 Introduction

In this paper they use information theory, in particular the rate-distortion theory (RDT) of lossy compression, to formulate a problem of encoding a variable (signal) so that the compressed representation contains the *relevant information* about some other variable.

Based on (Cover and Thomas 2006):

In the classical RDT we consider an information source generating sequences of i.i.d. random variables X from **a known probability distribution** $X \sim p(x)$ with a finite alphabet $X \in \mathcal{X}$. We encode the source sequence $x^n \in \mathcal{X}^n$ through an encoding function $f : \mathcal{X}^n \to \mathcal{Z}$ to an index $z \in \mathcal{Z} = \{1, 2, \ldots, 2^M\}$, where M is the number of bits we can use to represent the source sequence. The decoder maps the index $z \in \mathcal{Z}$ to a representation $\hat{x}^n \in \hat{\mathcal{X}}^n$ of the source sequence $x^n \in \mathcal{X}^n$ through a decoding function $g : \mathcal{Z} \to \hat{\mathcal{X}}^n$. The set of possible decodings $\{g(1), g(2), \ldots, g(2^M)\}$ constitutes the *codebook* and $f^{-1}(z)$ are the *assignment regions*. Replacement of X^n by $\hat{X}^n$ is commonly referred to as *vector quantization*, representation or reconstruction.

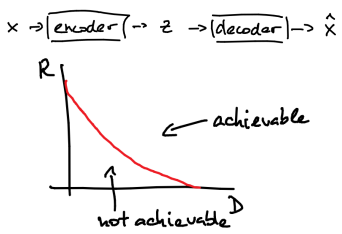

Figure 1: Rate-distortion encoder and decoder and $R(D)$ curve.

The RDT characterizes the tradeoff between the rate $R = M/n$ (average number of bits per symbol in the source sequence) and the distortion $D = \mathrm{E}_{p(x)} d(X^n, f(g(X^n)))$ associated with the code. The rate-distortion function $R(D)$ determines the smallest possible rate for a given distortion D.
I think: If $R = \mathrm{H}(X)$, this is a lossless autoencoding system where we can achieve zero distortion $D = 0$. With lower rates $R < \mathrm{H}(X)$ the distortion will increase.

Back to Tishby:

The problem in RDT is that it does not specify which distortion functi on $d(x, \hat{x})$ you shall use.[1] Once you pick it, it essentially determines which features of the signal will be considered as *relevant* and encoded into Z. However, without a definition of *relevance*, this is not a well posed problem.

What they propose here is to use an additional variable Y to determine what is *relevant* information. The structure of the problem is then: extract to $\hat{X}$ the info from X that is relevant for predicting Y. The choice of Y will determine the *relevant features* of the signal.

TLDR: Use info theory and rate-distortion theory of encoding $X \to Z \to \hat{X}$ to motivate encoding (compression $\approx$ feature extraction) strategy for classification problem $X \to Z \to \hat{X} \to \hat{Y}$. Careful, $Z \in 1, \ldots, 2^M$ is an index with M determining the rate of encoding $R = M/n$ (n is the length of the original sequence to be encoded).

1.2 Relevant quantization

Assume signal space $\mathcal{X}$ with fixed **known prob. distribution** $p(x)$ and $\hat{\mathcal{X}}$ its reconstruction space - quantized codebook. **Both $\mathcal{X}$ and $\hat{\mathcal{X}}$ spaces are finite $\approx$ discrete or quantized (if originally**

[1] Classical examples are the squared error loss or Hamming distance.

continuous).

For each $x \in \mathcal{X}$ we seek a possibly *stochastic* mapping to a codeword in the codebook $\hat{x} \in \hat{\mathcal{X}}$ with conditional prob. distribution $p(\hat{x} \mid x)$.

$$p(\hat{x}) = \sum_x p(x) p(\hat{x} \mid x) \quad . \tag{1.1}$$

Average volume of elements of $\mathcal{X}$ mapped to the same codeword $\hat{x} \in \hat{\mathcal{X}}$ is $2^{\mathrm{H}(X|\hat{X})}$

> This comes from the entropy $\mathrm{H}(X) = \mathrm{E} \log_2 1/p(x)$ being the optimal length of encoding. Think of uniform X with $|X| = c$ elements and $p(x) = 1/c$ which has entropy $\mathrm{H}(X) = 1/c \sum \log_2 c = \log_2 c$. Then volume is $|X| = c = 2^{\mathrm{H}(X)}$

What matters for the quality of the quantization $\mathcal{X}$ is a) the average number of bits per codeword $\hat{X} \in \hat{\mathcal{X}}$ and b) the expected distortion between the source and encoding $\mathrm{E}_{p(x)} d(x, \hat{x})$.

The average amount of information (in bits) we gain about X by knowing $\hat{X}$ (or equivalently, the reduction in the average number of bits per element in X knowing $\hat{X}$) is the mutual information

$$\mathrm{I}(X; \hat{X}) = \sum_{x \in \mathcal{X}} \sum_{\hat{x} \in \hat{\mathcal{X}}} p(x, \hat{x}) \log \frac{p(\hat{x} \mid x)}{p(\hat{x})} = \mathrm{H}(\hat{X}) - \mathrm{H}(\hat{X} \mid X) = \mathrm{H}(X) - \mathrm{H}(X \mid \hat{X}) \quad . \tag{1.2}$$

For stochastic mappings $\mathrm{H}(\hat{X} \mid X) \neq 0$ and $\mathrm{H}(\hat{X})$ is not what we want to minimize.

> **TLDR:** Treat source $\mathcal{X}$ and reconstruction $\hat{\mathcal{X}}$ spaces as discrete. Allow for stochastic mapping (encoding-decoding) $p(\hat{x}|x)$. Good reconstruction should have high mutual information with the source $\mathrm{I}(X; \hat{X}) = \sum_{x \in \mathcal{X}} \sum_{\hat{x} \in \hat{\mathcal{X}}} p(x, \hat{x}) \log \frac{p(\hat{x}|x)}{p(\hat{x})}$.

1.3 Relevance through distortion

The ability to reconstruct is measured in RDT by the distortion function $\mathcal{X} \times \hat{\mathcal{X}} \to \mathbb{R}_+$ with expected distortion

$$D = \mathrm{E}_{p(x,\hat{x})} d(x, \hat{x}) = \sum_{x \in \mathcal{X}} \sum_{\hat{x} \in \hat{\mathcal{X}}} d(x, \hat{x}) \tag{1.3}$$

which is presumed to be low for good reconstructions $\hat{X}$ and which implicitly specifies what are the most *relevant* aspects of X.

There is a monotonic trade-off between the rate R and the distortion D described by the rate-distortion function $R(D)$. The $R(D)$ function is defined as the minimal achievable rate (mutual information) with a given constraint D^* on the expected distortion

$$R(D) := \min_{p(\hat{x}|x) : \mathrm{E}_{p(x,\hat{x})} d(x,\hat{x}) \leq D^*} \mathrm{I}(X; \hat{X}) \tag{1.4}$$

or in the Lagrangian form as the minimization of

$$\mathcal{L}[p(\hat{x} \mid x)] = \mathrm{I}(X; \hat{X}) + \beta \mathrm{E}_{p(x,\hat{x})} d(x, \hat{x}) \quad , \tag{1.5}$$

where $\beta = -\frac{\mathrm{d}R}{\mathrm{d}D} > 0$

For each β (that is for each distortion constraint D^*) the $p(\hat{x} \mid x)$ can be found by *Blahut-Arimoto algorithm* which alternates between ensureing that $p(\hat{x}) = \sum_x p(x) p(\hat{x} \mid x)$ and minimizing the RDT

objective. It's important to note that it only finds the optimal partitioning $p(\hat{x} \mid x)$ over a given representation space $\hat{X}$. Finding optimal space $\hat{X}$ would need some sort of EM algorithm.

TLDR: In RDT we ere looking for $p(\hat{x} \mid x)$ which minimizes the mutual information $\mathrm{I}(X;\hat{X})$ (minimizes the rate $\approx$ maximizes the compression) with a limit on the expected distortion $\mathrm{E}_{p(x,\hat{x})} d(x,\hat{x}) \leq D^*$. This can be formulated as the minimization of

$$\mathcal{L}[p(\hat{x} \mid x)] = \mathrm{I}(X;\hat{X}) + \beta \mathrm{E}_{p(x,\hat{x})} d(x,\hat{x}) \ . \tag{1.6}$$

1.4 Relevance through other variable - information bottleneck

As before, we want the quantization $\hat{X}$ to compress X as much as possible but instead of looking at distortion $d(x,\hat{x})$, we will look at how much information about some third variable Y the quantization $\hat{X}$ can capture $\mathrm{I}(\hat{X},Y)$. We assume that the original signal has positive mutual information with the variable $\mathrm{I}(X;Y)$ and that the **true joint distribution** $p(x,y)$ **is known.**

As lossy compression cannot convey more information than the original data we have $\mathrm{I}(\hat{X};Y) \leq \mathrm{I}(X;Y)$. The trade-off we look for now is between the rate $\mathrm{I}(X;\hat{X})$ (compression) while preserving meaningful info about Y: we pass the information X has about Y through a *bottleneck* representation $\hat{X}$.

Similarly to before, we find the optimal $p(\hat{x} \mid x)$ by minimizing the functional

$$\mathcal{L}[p(\hat{x} \mid x)] = \mathrm{I}(X;\hat{X}) - \beta \mathrm{I}(\hat{X};Y) \ , \tag{1.7}$$

where instead of minimizing the expected distortion $\mathrm{E}_{p(x,\hat{x})} d(x,\hat{x})$ we maximize the mutual information $\mathrm{I}(\hat{X};Y)$.

They then proof using the machinery of the Blahut-Arimoto algorithm that the optimal solution is

$$p(\hat{x} \mid x) = \frac{p(\hat{x})}{Z(x,\beta)} \exp\left[-\beta \, \mathrm{KL}\left(p(y \mid x) \,\|\, p(y \mid \hat{x})\right)\right] \ , \tag{1.8}$$

where $Z(x,\beta)$ is the normalization constant (nothing to do with the encoder $\mathcal{Z}$.) This suggest the KL divergence is the *correct* distortion measure in this seeting. They further propose an algorithm which is alternating between optimizing $p(\hat{x} \mid x)$, and making sure $p(\hat{x})$ and $p(y \mid \hat{x})$ are consistent with it and the known $p(x,y)$.

TLDR: Use some other variable Y with $\mathrm{I}(X;Y) > 0$ to guide the quality of the representation $\hat{X}$. It should still compress (with low $\mathrm{I}(X;\hat{X})$) but instead of $d(x,\hat{x})$ use $\mathrm{I}(\hat{X};Y)$ to see how much the representation preserves from Y. The optimization is

$$\mathcal{L}[p(\hat{x} \mid x)] = \mathrm{I}(X;\hat{X}) - \beta \mathrm{I}(\hat{X};Y) \ , \tag{1.9}$$

and the solution is alternating algo similar to Arimoto-Blahut.

Concluding remarks: It is not clear to me what (if anything) can break if we move from the discrete case considered here to the continuous case. But, more importantly, the assumption that we have access to the true distribution $p(x,y)$ certainly does not hold in ML.

2 Tishby's information bottleneck principle in DL

Paper: Naftali Tishby and Noga Zaslavsky. "Deep Learning and the Information Bottleneck Principle". In: *IEEE Information Theory Workshop (ITW)*. 2015

Notes taken: 25/12/2019

2.1 Introduction

They formulate the goal of deep learning as an information theoric trade-off between compression and prediction (building on the information bottleneck method 1) and give some guarantees for generalization.

2.1.1 The DNN setting

They consider MLPs with sigmoid neurons, high dimensional inputs X and classification (multi-class) output Y. The optimisation is performed by SGD through backpropagation of some loss.

2.1.2 The information bottleneck principle

The principle suggests to encode X into $\hat{X}$ such that it squeezes out of X all information not relevant for predicting Y. $\hat{X}$ can be seen as the minimal sufficient statistic of X with respect to Y, ot the simplest representation of X that captures the mutual info $\mathrm{I}(X;Y)$. For the classification case they assume a Markov chain $Y \to X \to \hat{X}$ with conditional independence assumption $Y \perp \hat{X} \mid X$ described by the joint density $p(y,x,\hat{x}) = p(y)p(x \mid y)\,p(\hat{x} \mid x)$ and the data processing inequality (DPI) $I(Y;X) \geq I(Y;\hat{X})$.

The information bottleneck (IB) method proposes to learn the stochastic mapping $p(\hat{x} \mid x)$ by minimizing

$$\mathcal{L}[p(\hat{x} \mid x)] = \mathrm{I}(X;\hat{X}) - \beta \mathrm{I}(\hat{X};Y) \ . \tag{2.1}$$

They suggest to reformulate it reformulate it as the minimization

$$\mathcal{L}[p(\hat{x} \mid x)] = \mathrm{I}(X;\hat{X}) + \beta \mathrm{I}(X;Y \mid \hat{X}) \ , \tag{2.2}$$

that is minimizing the residual information about Y in X not captured by $\hat{X}$.

2.1.3 IB for DNN

They see the DNN layers as forming a Markov chain $Y \to X \to h_1 \to \ldots \to h_n \to \hat{Y}$ in which the DPI holds $\mathrm{I}(Y;X) \geq \mathrm{I}(Y;h_1) \geq \ldots \geq \mathrm{I}(Y;h_n) \geq \mathrm{I}(Y;\hat{Y})$, where equality holds only if each layer is sufficient statistic of its input with respect to Y.

From learning perspective, each layer should try to minimize $\mathrm{I}(h_{i-1};h_i)$ while maximizing $\mathrm{I}(Y;h_i)$. They have got some bounds on the prediction error based on $\mathrm{I}(Y;h_i)$. Also the optimal theoretical limit of equation (2.2) can be evaluated for each intermediate layer conditioning on the previous layers. Each consecutive layer compresses the inputs increasing the distortion.

2.1.4 Generalization bounds

The IB curve obtained by minimizing (2.2) is a property of a joint distribution $p(x,y)$ which is unknown in ML. Assuming bouinded cardinality of the representation $K = |\hat{X}|$ (that is of the layers h_i) the difference between the true mutual information $\mathrm{I}(X;\hat{X})$ resp. $\mathrm{I}(\hat{X};Y)$ and their sample estimates increases with K but not with the cardinality of X. Which means the IB curve can be well estimated for compressed representations but badly estimated for complex representations. Using these bounds, they sketch worst-case bound on the out-of-sample RD curve which actually has an optimal point with minimal distortion and corresponding rate. They claim that as the compression progresses we move closer in the information plane to this optimal point.

TLDR: Transfer the information bottleneck method to DNN classification learning. For compressed representation $\hat{X}$ of the signal X and the target Y the optimization problem is

$$\mathcal{L}[p(\hat{x} \mid x)] = \mathrm{I}(X;\hat{X}) + \beta \mathrm{I}(X;Y \mid \hat{X}) \ . \tag{2.3}$$

In the DNN case each hidden layer h_i is seen as the compressed representation $\hat{X}$ so that we have the following DPI $\mathrm{I}(Y;X) \geq \mathrm{I}(Y;h_1) \geq \ldots \geq \mathrm{I}(Y;h_n) \geq \mathrm{I}(Y;\hat{Y})$. They also have some generalization bounds claiming that what matters is the cardinality of the representations $K = |\hat{X}|$ but not of $|X|$.

Concluding remarks: Rather straightforward extension of IB method to DNNs. I'm not sure how reasonable is the generalization analysis (haven't read the other paper they wrote and cite) but it seems to somehow rely on the discrete nature of the source and encoding spaces which is gnerally not true in DL.

3 Tishby's information theory of DL

Video: Naftali Tishby. *Information Theory of Deep Learning - Naftali Tishby*. Yandex, 2017
Notes taken: 26/12/2019

Bringing information theory into classification problems solved by DL.

Transferring information bottleneck method and data processing inequality (DPI) into the learning. The DPI basically says that for any three random variables in a Markov chain Y->X->H we have $\mathrm{I}(Y,X) \geq \mathrm{I}(Y,H)$. Thinking about hidden layers H as useful compression, we should try to compress H to reduce as much as possible $\mathrm{I}(X,H)$ while keeping $\mathrm{I}(H,Y) \approx$ information bottleneck method.

He is very keen to replace the standard PAC learning theory by his IB based theory so he is well aware of the problem of generalization. But I don't find in his talk any really convincing arguments. He simplifies so much that in the end it seem he is missing the point.

He has plenty of stories about how what we see in training is first a drift of the layers to remember the data followed by a stage of compression. It all seems to me like a somewhat far-fetched try to push the coding theory onto the learning. But for it to work you would need to shift the DL training from *continuous to discrete* problem and instead of the nets being *deterministic*, have them as *stochastic*.

It seems to me that plenty of what he shows is rather the *result of his evaluation procedure*, e.g. the binning which suggests loss of information while in the continuous case no info is lost and some sort of clustering which again never really happens and rather the Bayes decision rule is applied over the continuous outputs.

Other questions: As he says, each of the layers is just a deterministic transformation of the previous layer which is moreover in the case of simple nonlinearities such as sigmoid or tanh invertible. He also says that mutual information is invariant under invertible transformations $\mathrm{I}(X,Y) = \mathrm{I}(f(X),g(Y))$. If both of these are true, how could any information be lost?

He also admits that the learning rule cannot be completely deterministic because otherwise his theory breaks and that they kind of get away from it by the binning. But this is not what really happens. What in his talk is real effect and what just evaluation artifact?

He argues by typicality that the network actually learns some clusters. I can see the network learns clusters but the typicality arguments seem far-fetched.

The usual coding theory is based on knowing the true $p(x,y)$ distribution but we only observe samples of it. He admits this is a problem but almost trivializes this too much.

He has a graph showing how during the training gradients first have high mean and small std (across the batches within epoch) and some point shift to low means and high gradients. I find this perfectly natural. The network first needs to learn so all gradients are high and not too different as all the updates from all the batches push the weights to something reasonable. After this happens, the gradients become smaller in mean (reaching a plateau so don't need so big changes) but each batch suggests a move in different direction. In my head this does not necessarily have to do anything with compression, just the way the optimization happens.

4 Saxe (Harvard): IB in DL - critique of Tishby's findings

Paper: Andrew Michael Saxe et al. "On the Information Bottleneck Theory of Deep Learning". In: *ICLR*. 2018

Notes taken: 26/12/2019

They question the conclusions of (Tishby 2017) (see section 3) who claims that a) DL training consists of two phases - fitting followed by compression, b) it is due to the compression phase that DN don't overfit, c) the compression appears due to diffusion-like effect of SGD. Here they show that none of these is generally true (mainly empirically with some theory behind but not too elaborate). The compression that Tishby observed is more due to nonlinearity they used (tanh) which has a double-sided saturation effect (unlike ReLu which is unbounded on one side). They also show that you can have good generalization without compression and that compression does not arise from stochasticity of SGD but can arise in batch as well.

There is a discussion related to the evaluation of the mutual information which usually calculates the info and entropy metrics as if the variables were discrete (through binning) but since they are not, this is just an arbitrary bias introduced to the evaluation.

Also, for deterministic functions f the mutual information $\mathrm{I}(X, h = f(X))$ is infinity, it is in general impossible to analyse mutual info of the deterministic NN layers without introducing some error.

Concluding remarks: The paper caused quite a lot of controversy around the IB method by criticising initial Tishby's findings. Though they often present similarly flawed results (little theory, basic experiments) they pose the right questions and open room for discussion.

5 Goldfeld (MIT): Estimating Info Flow in DNNs

Paper: Ziv Goldfeld et al. "Estimating Information Flow in Deep Neural Networks". In: *ICML*. 2019

Notes taken: 26/12/2019

Coming back to recent papers of Shwartz-Ziv and Tishby 2017; Saxe et al. 2018 they recognize that using mutual information I(.;.) for measuring the effect of learning does not make sense in non-stochastic setting because it is either constant (in the discrete setting) or infinite (in the continuous). Furthermore, the compression effect empirically observed in the previous papers is likely an artefact of the estimation procedure of I(.;.) relying on binning the neuron output space.

To address this they propose a stochastic DNN framework where each neuron output is contaminated by Gaussian noise. This makes I(.;.) express something reasonable *(though to me it seems just making the problem even more contrived because what we will be measuring is somehow the resiliency to the imputed noise)*. I(.;.) still has no exact analytical expression due to the complex transformations of the variables (and the distributions) in the DNN.

They propose an estimator of I(.;.) (based on their previous work) which uses the empirical estimates of the differential entropies of the Gaussian noisy signals (Gaussian noisy channel in the info theory jargon). These rely on the fact that the data sample is i.i.d. and therefore entropy estimates should converge to the true differential entropies with enough samples. They still need some Monte Carlo integration to really evaluate the entropies but they don't elaborate. Obviously, the number of samples is critical and they do admit it but provide results for manageable sample sizes claiming they should be good enough.

Finally, they empirically explore the behaviour of the mutual information $\mathrm{I}(X;T_k)$ and the entropy $\mathrm{H}(binned(T_k))$, where T_k is the layer output after adding the noise. They conclude that what is actually happening during the training across the layers is better and better clustering to finally correspond to the output categories. The compression previously observed was actually this clustering which is naturally captured in the bin-based entropy but not in the true mutual information, and certainly not in the deterministic case when measuring mutual info makes no sense. Furthermore, compression measured through mutual information does not seem to causally relate to generalization. The clustering effect of the layers (its geometry) is worth further study.

TLDR: Mutual information is useless for tracing learning through layers of deterministic nets (it's constant for discrete and infinite for continuous). What has been previously observed as compression is artefact of estimating differential entropy via binning. The binning estimates clustering within the transformed space and it is the clustering effect which is important for accuracy and generalization (not the compression measured by mutual information). Whatever the previous papers were saying about compression holds if you **replace compression with clustering**.

Concluding remarks: To me the clustering effect of the neurons and layers seems so obvious that I find it difficult to believe no one discussed it before. However, if you think about clustering it is a form of compression so one needs to be careful about the terminology. Perhaps rather than clustering, what is happening is some sort of concentration, mode creation in the distributions? Well, this is sort of clustering as well (aka Gaussian mixture).

6 Slava's WP linking VAEs and GANs through information bottleneck

Paper: Slava Voloshynovskiy et al. "Information Bottleneck through Variational Glasses". In: *arXiv:1912.00830 [cs]* (2019)

Notes taken: 26/12/2019

Warning: In this summary I deviate quite a bit from the original paper because the original structure and notation is difficult to follow.

6.1 Introduction

They aim at transferring the IB method from supervised learning to unsupervised learning of the VAE and GAN family and try to link all these together by formulating unified learning objectives. This is achieved through various decomposition and reformulations of the original IB method (not really achieved for GANs, in my view.)

Thought: The IB supervised method was initially motivated by the unsupervised rate-distortion coding. Isn't this a bit running in circles?

6.2 IB for supervised models

Standard supervised classification setup with observations $\{(\mathbf{x}_i, \mathbf{c}_i) \in \mathbb{R}^d \times \mathcal{M}\}_{i=1}^N$ assuming true generative process $p(\mathbf{x}, \mathbf{c}) = p(\mathbf{c})p(\mathbf{x} \mid \mathbf{c})$.

According to Tishby's bottleneck method (Tishby, Pereira, et al. 2000; Tishby and Zaslavsky 2015) (see sections 1, 2) the optimization problem is the minimization of the functional

$$\begin{aligned}\mathcal{L}^S[q_\phi(\mathbf{z} \mid \mathbf{x})] &:= \mathrm{I}_\phi(\mathbf{X};\mathbf{Z}) - \beta \mathrm{I}(\mathbf{Z};\mathbf{C}) \\ &= \mathrm{H}_\phi(\mathbf{Z}) - \mathrm{H}_\phi(\mathbf{Z} \mid \mathbf{X}) - \beta \Big(\mathrm{H}(\mathbf{C}) - \mathrm{H}_\phi(\mathbf{C} \mid \mathbf{Z})\Big) \ , \end{aligned} \tag{6.1}$$

where $\mathbf{Z}$ is a representation of $\mathbf{X}$ with a learned probabilistic mapping $q_\phi(\mathbf{z} \mid \mathbf{x})$ and Markov chain condition $\mathbf{C} \to \mathbf{X} \to \mathbf{Z}$ so that $(\mathbf{Z} \perp \mathbf{C}) \mid \mathbf{X}$.

In the original Tishby's paper it was assumed that the true generative distribution $p(\mathbf{x}, \mathbf{c})$ and the support sets of $\mathbf{X}$ and $\mathbf{Z}$ ***are known*** and they are both finite (i.e. discrete random vars). However, here they use integrals which means they assume continuous $\mathbf{X}$ and $\mathbf{Z}$ which corresponds better to DL (but I'm not sure what are the consequences for the mutual info which breaks for deterministic mappings).

The following relations are needed for the entropy evaluations in 6.1) hold

$$p(\mathbf{x}, \mathbf{c}) = p(\mathbf{c})p(\mathbf{x} \mid \mathbf{c}) \qquad p(\mathbf{c} \mid \mathbf{x}) = \frac{p(\mathbf{x}, \mathbf{c})}{p(\mathbf{x})}$$

$$p_\phi(\mathbf{x}, \mathbf{z}) = p(\mathbf{x})q_\phi(\mathbf{z} \mid \mathbf{x}) \qquad p_\phi(\mathbf{z}) = \int_{\mathbf{x}} p(\mathbf{x})q_\phi(\mathbf{z} \mid \mathbf{x}) \, \mathrm{d}\mathbf{x} \qquad p_\phi(\mathbf{x} \mid \mathbf{z}) = \frac{p(\mathbf{x})q_\phi(\mathbf{z} \mid \mathbf{x})}{\int_{\mathbf{x}} p(\mathbf{x})q_\phi(\mathbf{z} \mid \mathbf{x}) \, \mathrm{d}\mathbf{x}}$$

$$p_\phi(\mathbf{c}, \mathbf{z}) = \int_{\mathbf{x}} p(\mathbf{c} \mid \mathbf{x}) p_\phi(\mathbf{x}, \mathbf{z}) \, \mathrm{d}\mathbf{x} = \int_{\mathbf{x}} \frac{p(\mathbf{x}, \mathbf{c})}{p(\mathbf{x})} p(\mathbf{x})q_\phi(\mathbf{z} \mid \mathbf{x}) \, \mathrm{d}\mathbf{x} = \int_{\mathbf{x}} p(\mathbf{x}, \mathbf{c})q_\phi(\mathbf{z} \mid \mathbf{x}) \, \mathrm{d}\mathbf{x}$$

$$p_\phi(\mathbf{c} \mid \mathbf{z}) = \int_{\mathbf{x}} p(\mathbf{c} \mid \mathbf{x}) p_\phi(\mathbf{x} \mid \mathbf{z}) \, \mathrm{d}\mathbf{x} = \int_{\mathbf{x}} \frac{p(\mathbf{x}, \mathbf{c})}{p(\mathbf{x})} \frac{p(\mathbf{x})q_\phi(\mathbf{z} \mid \mathbf{x})}{\int_{\mathbf{x}} p(\mathbf{x})q_\phi(\mathbf{z} \mid \mathbf{x})\mathbf{x}} \, \mathrm{d}\mathbf{x} = \int_{\mathbf{x}} \frac{p(\mathbf{x}, \mathbf{c})q_\phi(\mathbf{z} \mid \mathbf{x})}{\int_{\mathbf{x}} p(\mathbf{x})q_\phi(\mathbf{z} \mid \mathbf{x}) \, \mathrm{d}\mathbf{x}} \, \mathrm{d}\mathbf{x}$$

$$\mathrm{H}_\phi(\mathbf{Z}) = -\int_{\mathbf{z}} p_\phi(\mathbf{z}) \log p_\phi(\mathbf{z})\, \mathrm{d}\mathbf{z} = -\int_{\mathbf{z}}\int_{\mathbf{x}} p(\mathbf{x}) q_\phi(\mathbf{z} \mid \mathbf{x}) \log\left(\int_{\mathbf{x}} p(\mathbf{x}) q_\phi(\mathbf{z} \mid \mathbf{x})\, \mathrm{d}\mathbf{x}\right) \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{z}$$

$$\mathrm{H}_\phi(\mathbf{Z} \mid \mathbf{X}) = -\int_{\mathbf{x}}\int_{\mathbf{z}} p(\mathbf{x}) q_\phi(\mathbf{z} \mid \mathbf{x}) \log q_\phi(\mathbf{z} \mid \mathbf{x})\, \mathrm{d}\mathbf{z}\, \mathrm{d}\mathbf{x} \qquad \mathrm{H}(\mathbf{C}) = -\sum_{\mathbf{c}} p(\mathbf{c}) \log p(\mathbf{c})$$

$$\mathrm{H}_\phi(\mathbf{C} \mid \mathbf{Z}) = -\int_{\mathbf{z}} \sum_{\mathbf{c}} p_\phi(\mathbf{z}) p_\phi(\mathbf{c} \mid \mathbf{z}) \log p_\phi(\mathbf{c} \mid \mathbf{z})\, \mathrm{d}\mathbf{z} = -\int_{\mathbf{z}}\int_{\mathbf{x}} \sum_{\mathbf{c}} p(\mathbf{x}, \mathbf{c}) q_\phi(\mathbf{z} \mid \mathbf{x}) \log p_\phi(\mathbf{c} \mid \mathbf{z})\, \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{z}$$

$$= -\int_{\mathbf{z}}\int_{\mathbf{x}} \sum_{\mathbf{c}} p(\mathbf{x}, \mathbf{c}) q_\phi(\mathbf{z} \mid \mathbf{x}) \log\left(\int_{\mathbf{x}} \frac{p(\mathbf{x}, \mathbf{c}) q_\phi(\mathbf{z} \mid \mathbf{x})}{\int_{\mathbf{x}} p(\mathbf{x}) q_\phi(\mathbf{z} \mid \mathbf{x})\, \mathrm{d}\mathbf{x}}\, \mathrm{d}\mathbf{x}\right) \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{z}$$

For the moment, we cannot evaluate any of the above entropies and learn $q_\phi(\mathbf{z} \mid \mathbf{x})$ because a) the true generative distribution is unknown, b) integrations over the $\mathbf{X}$ and $\mathbf{Z}$ spaces.

In the next step they introduce a learnable approximation to the classification distribution $p_\phi(\mathbf{c} \mid \mathbf{z}) \approx p_\theta(\mathbf{c} \mid \mathbf{z})$ parametrized by θ (they don't really explain why) with a cross-entropy $\mathrm{H}_{\phi,\theta}(\mathbf{C} \mid \mathbf{Z})$

$$\mathrm{H}_\phi(\mathbf{C} \mid \mathbf{Z}) \le \mathrm{H}_{\phi,\theta}(\mathbf{C} \mid \mathbf{Z}) = -\int_{\mathbf{z}}\int_{\mathbf{x}} \sum_{\mathbf{c}} p(\mathbf{x}, \mathbf{c}) q_\phi(\mathbf{z} \mid \mathbf{x}) \log p_\theta(\mathbf{c} \mid \mathbf{z})\, \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{z}$$

$$\mathrm{H}_{\phi,\theta}(\mathbf{C} \mid \mathbf{Z}) - \mathrm{H}_\phi(\mathbf{C} \mid \mathbf{Z}) = \int_{\mathbf{z}} p_\phi(\mathbf{z})\, \mathrm{KL}\left(p_\phi(\mathbf{c} \mid \mathbf{z}) \,\|\, p_\theta(\mathbf{c} \mid \mathbf{z})\right) \mathrm{d}\mathbf{z}\ , \tag{6.2}$$

where the inequality holds by standard properties of cross-entropy and the positive KL.

They replace the classification entropy with the cross-entropy in the objective function 6.1 to obtain an upper bound which can be minimized instead

$$\mathcal{L}^S[q_\phi(\mathbf{z} \mid \mathbf{x})] \le \mathcal{L}^{SB}[q_\phi(\mathbf{z} \mid \mathbf{x}), p_\theta(\mathbf{c} \mid \mathbf{z})] := \mathrm{H}_\phi(\mathbf{Z}) - \mathrm{H}_\phi(\mathbf{Z} \mid \mathbf{X}) + \beta \mathrm{H}_{\phi,\theta}(\mathbf{C} \mid \mathbf{Z}) - \beta \mathrm{H}(\mathbf{C}). \tag{6.3}$$

Note that the last term is constant (though unknown) and can be dropped from the optimization objective. Also note that for any pair $(q_\phi(\mathbf{z} \mid \mathbf{x}), p_\theta(\mathbf{c} \mid \mathbf{z}))$ (including the optimal) we have

$$\mathcal{L}^{SB}[q_\phi(\mathbf{z} \mid \mathbf{x}), p_\theta(\mathbf{c} \mid \mathbf{z})] = \mathcal{L}^S[q_\phi(\mathbf{z} \mid \mathbf{x})] + \int_{\mathbf{z}} p_\phi(\mathbf{z})\, \mathrm{KL}\left(p_\phi(\mathbf{c} \mid \mathbf{z}) \,\|\, p_\theta(\mathbf{c} \mid \mathbf{z})\right) \mathrm{d}\mathbf{z} \tag{6.4}$$

TLDR: Start from Tishby's IB method objective for classification problem. In addition to learning the ϕ parameter of the representation stochastic mapping $q_\phi(\mathbf{z} \mid \mathbf{x})$, learn also the θ parameters of the variational approximation to the classification entropy $p_\theta(\mathbf{c} \mid \mathbf{z}) \approx p_\phi(\mathbf{c} \mid \mathbf{z})$. Instead of optimising the original objective $\mathcal{L}^S[q_\phi(\mathbf{z} \mid \mathbf{x})]$, minimise its upper bound

$$\mathcal{L}^{SB}[q_\phi(\mathbf{z} \mid \mathbf{x}), p_\theta(\mathbf{c} \mid \mathbf{z})] = \mathcal{L}^S[q_\phi(\mathbf{z} \mid \mathbf{x})] + \int_{\mathbf{z}} p_\phi(\mathbf{z})\, \mathrm{KL}\left(p_\phi(\mathbf{c} \mid \mathbf{z}) \,\|\, p_\theta(\mathbf{c} \mid \mathbf{z})\right) \mathrm{d}\mathbf{z}$$
$$= \mathrm{H}_\phi(\mathbf{Z}) - \mathrm{H}_\phi(\mathbf{Z} \mid \mathbf{X}) + \beta \mathrm{H}_{\phi,\theta}(\mathbf{C} \mid \mathbf{Z})\ ,$$

where $\mathrm{H}_{\phi,\theta}(\mathbf{C} \mid \mathbf{Z}) = -\int_{\mathbf{z}}\int_{\mathbf{x}} \sum_{\mathbf{c}} p(\mathbf{x}, \mathbf{c}) q_\phi(\mathbf{z} \mid \mathbf{x}) \log p_\theta(\mathbf{c} \mid \mathbf{z})\, \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{z}$ is the cross-entropy for ecoding $p_\theta(\mathbf{c} \mid \mathbf{z})$ via $p_\phi(\mathbf{c} \mid \mathbf{z})$. It is not clear how this problem can be optimised.

6.3 Information bottleneck for unsupervised problem

The set-up changes in that we only observe samples of $\mathbf{X} \sim p(\mathbf{x})$ and want to predict $\mathbf{X}$ from the representation $\mathbf{Z}$.

Essentially, they replace $\mathbf{C}$ with $\mathbf{X}$ in all the above.

The original objective is

$$\begin{aligned}\mathcal{L}^U[q_\phi(\mathbf{z}\mid\mathbf{x})] &:= \mathrm{I}_\phi(\mathbf{X};\mathbf{Z}) - \beta\mathrm{I}(\mathbf{Z};\mathbf{X}) \\ &= \mathrm{H}_\phi(\mathbf{Z}) - \mathrm{H}_\phi(\mathbf{Z}\mid\mathbf{X}) - \beta\Big(\mathrm{H}(\mathbf{X}) - \mathrm{H}_\phi(\mathbf{X}\mid\mathbf{Z})\Big)\ ,\end{aligned} \tag{6.5}$$

with the following relations

$$p_\phi(\mathbf{x},\mathbf{z}) = p(\mathbf{x})q_\phi(\mathbf{z}\mid\mathbf{x}) \qquad p_\phi(\mathbf{z}) = \int_{\mathbf{x}} p(\mathbf{x})q_\phi(\mathbf{z}\mid\mathbf{x})\,\mathrm{d}\,\mathbf{x} \qquad p_\phi(\mathbf{x}\mid\mathbf{z}) = \frac{p(\mathbf{x})q_\phi(\mathbf{z}\mid\mathbf{x})}{\int_{\mathbf{x}} p(\mathbf{x})q_\phi(\mathbf{z}\mid\mathbf{x})\,\mathrm{d}\,\mathbf{x}}$$

$$\mathrm{H}_\phi(\mathbf{Z}) = -\int_{\mathbf{z}} p_\phi(\mathbf{z})\log p_\phi(\mathbf{z})\,\mathrm{d}\,\mathbf{z} = -\int_{\mathbf{z}}\int_{\mathbf{x}} p(\mathbf{x})q_\phi(\mathbf{z}\mid\mathbf{x})\log\left(\int_{\mathbf{x}} p(\mathbf{x})q_\phi(\mathbf{z}\mid\mathbf{x})\,\mathrm{d}\,\mathbf{x}\right)\mathrm{d}\,\mathbf{x}\,\mathrm{d}\,\mathbf{z}$$

$$\mathrm{H}_\phi(\mathbf{Z}\mid\mathbf{X}) = -\int_{\mathbf{x}}\int_{\mathbf{z}} p(\mathbf{x})q_\phi(\mathbf{z}\mid\mathbf{x})\log q_\phi(\mathbf{z}\mid\mathbf{x})\,\mathrm{d}\,\mathbf{z}\,\mathrm{d}\,\mathbf{x} \qquad \mathrm{H}(\mathbf{X}) = -\int_{\mathbf{x}} p(\mathbf{x})\log p(\mathbf{x})$$

$$\begin{aligned}\mathrm{H}_\phi(\mathbf{X}\mid\mathbf{Z}) &= -\int_{\mathbf{z}}\int_{\mathbf{x}} p_\phi(\mathbf{z})p_\phi(\mathbf{x}\mid\mathbf{z})\log p_\phi(\mathbf{x}\mid\mathbf{z})\,\mathrm{d}\,\mathbf{z} = -\int_{\mathbf{z}}\int_{\mathbf{x}} p(\mathbf{x})q_\phi(\mathbf{z}\mid\mathbf{x})\log p_\phi(\mathbf{x}\mid\mathbf{z})\,\mathrm{d}\,\mathbf{x}\,\mathrm{d}\,\mathbf{z} \\ &= -\int_{\mathbf{z}}\int_{\mathbf{x}} p(\mathbf{x})q_\phi(\mathbf{z}\mid\mathbf{x})\log\left(\frac{p(\mathbf{x},\mathbf{c})q_\phi(\mathbf{z}\mid\mathbf{x})}{\int_{\mathbf{x}} p(\mathbf{x})q_\phi(\mathbf{z}\mid\mathbf{x})\,\mathrm{d}\,\mathbf{x}}\right)\mathrm{d}\,\mathbf{x}\,\mathrm{d}\,\mathbf{z}\end{aligned}$$

Next they introduce the learnable approximation $p_\theta(\mathbf{x}\mid\mathbf{z}) \approx p_\phi(\mathbf{x}\mid\mathbf{z})$ parametrized by θ with the cross-entropy $\mathrm{H}_{\phi,\theta}(\mathbf{X}\mid\mathbf{Z})$

$$\begin{aligned}\mathrm{H}_\phi(\mathbf{X}\mid\mathbf{Z}) \le \mathrm{H}_{\phi,\theta}(\mathbf{X}\mid\mathbf{Z}) &= -\int_{\mathbf{z}}\int_{\mathbf{x}} p(\mathbf{x})q_\phi(\mathbf{z}\mid\mathbf{x})\log p_\theta(\mathbf{x}\mid\mathbf{z})\,\mathrm{d}\,\mathbf{x}\,\mathrm{d}\,\mathbf{z} \\ \mathrm{H}_{\phi,\theta}(\mathbf{X}\mid\mathbf{Z}) - \mathrm{H}_\phi(\mathbf{X}\mid\mathbf{Z}) &= \int_{\mathbf{z}} p_\phi(\mathbf{z})\,\mathrm{KL}\Big(p_\phi(\mathbf{x}\mid\mathbf{z})\,\|\,p_\theta(\mathbf{x}\mid\mathbf{z})\Big)\,\mathrm{d}\,\mathbf{z}\ .\end{aligned} \tag{6.6}$$

Finally, they plug this into the objective so that

$$\mathcal{L}^U[q_\phi(\mathbf{z}\mid\mathbf{x})] \le \mathcal{L}^{UB}[q_\phi(\mathbf{z}\mid\mathbf{x}), p_\theta(\mathbf{x}\mid\mathbf{z})] := \mathrm{H}_\phi(\mathbf{Z}) - \mathrm{H}_\phi(\mathbf{Z}\mid\mathbf{X}) + \beta\mathrm{H}_{\phi,\theta}(\mathbf{X}\mid\mathbf{Z}) - \beta\mathrm{H}(\mathbf{X}). \tag{6.7}$$

The last constant (though unknown) term can again be dropped and we also have the bound relation for any pair $(q_\phi(\mathbf{z}\mid\mathbf{x}), p_\theta(\mathbf{x}\mid\mathbf{z}))$ (including the optimal)

$$\mathcal{L}^{UB}[q_\phi(\mathbf{z}\mid\mathbf{x}), p_\theta(\mathbf{x}\mid\mathbf{z}) = \mathcal{L}^U[q_\phi(\mathbf{z}\mid\mathbf{x})] + \int_{\mathbf{z}} p_\phi(\mathbf{z})\,\mathrm{KL}\Big(p_\phi(\mathbf{x}\mid\mathbf{z})\,\|\,p_\theta(\mathbf{x}\mid\mathbf{z})\Big)\,\mathrm{d}\,\mathbf{z} \tag{6.8}$$

TLDR: In the unsupervised setting we only have $\mathbf{X} \sim p(\mathbf{x})$ but will use exactly the same machinery for predicting (reconstructing) $\mathbf{X}$ from $\mathbf{Z}$ as in the supervised setting simply by replacing $\mathbf{C}$ with $\mathbf{X}$ everywhere. We learn the distribution $q_\phi(\mathbf{z}\mid\mathbf{x})$ and the variational approximation $p_\theta(\mathbf{x}\mid\mathbf{z}) \approx p_\phi(\mathbf{x}\mid\mathbf{z})$ by minimising

$$\begin{aligned}\mathcal{L}^{UB}[q_\phi(\mathbf{z}\mid\mathbf{x}), p_\theta(\mathbf{x}\mid\mathbf{z})] &= \mathcal{L}^U[q_\phi(\mathbf{z}\mid\mathbf{x})] + \int_{\mathbf{z}} p_\phi(\mathbf{z})\,\mathrm{KL}\Big(p_\phi(\mathbf{x}\mid\mathbf{z})\,\|\,p_\theta(\mathbf{x}\mid\mathbf{z})\Big)\,\mathrm{d}\,\mathbf{z} \\ &= \mathrm{H}_\phi(\mathbf{Z}) - \mathrm{H}_\phi(\mathbf{Z}\mid\mathbf{X}) + \beta\mathrm{H}_{\phi,\theta}(\mathbf{X}\mid\mathbf{Z})\ ,\end{aligned}$$

with $\mathrm{H}_{\phi,\theta}(\mathbf{X}\mid\mathbf{Z}) = -\int_{\mathbf{z}}\int_{\mathbf{x}} p(\mathbf{x})q_\phi(\mathbf{z}\mid\mathbf{x})\log p_\theta(\mathbf{x}\mid\mathbf{z})\,\mathrm{d}\,\mathbf{x}\,\mathrm{d}\,\mathbf{z}$.

6.3.1 Rewriting the objective

We can rewrite the final objective 6.7 as

$$\begin{aligned}\mathcal{L}^{UB}[q_\phi(\mathbf{z}\mid\mathbf{x}), p_\theta(\mathbf{x}\mid\mathbf{z})] = &-\int_{\mathbf{z}}\int_{\mathbf{x}} p(\mathbf{x}) q_\phi(\mathbf{z}\mid\mathbf{x}) \log \frac{p_\phi(\mathbf{z})}{q_\phi(\mathbf{z}\mid\mathbf{x})}\,\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{z} \\ &-\beta\int_{\mathbf{z}}\int_{\mathbf{x}} p(\mathbf{x}) q_\phi(\mathbf{z}\mid\mathbf{x}) \log p_\theta(\mathbf{x}\mid\mathbf{z})\,\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{z} \end{aligned} \tag{6.9}$$

They introduce another variational approximation $p_\xi(\mathbf{z}) \approx p_\phi(\mathbf{z})$, this time of the marginal of $\mathbf{Z}$, and bring it into the objective

$$\begin{aligned}\mathcal{L}^{UB}[q_\phi(\mathbf{z}\mid\mathbf{x}), p_\theta(\mathbf{x}\mid\mathbf{z}), p_\xi(\mathbf{z})] &= -\mathbb{E}_{p(\mathbf{x})}\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \log \frac{p_\phi(\mathbf{z}) p_\xi(\mathbf{z})}{p_\xi(\mathbf{z}) q_\phi(\mathbf{z}\mid\mathbf{x})} - \beta\mathbb{E}_{p(\mathbf{x})}\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \log p_\theta(\mathbf{x}\mid\mathbf{z}) \\ &= \mathbb{E}_{p(\mathbf{x})}\mathrm{KL}\left(q_\phi(\mathbf{z}\mid\mathbf{x}) \,\|\, p_\xi(\mathbf{z})\right) - \mathrm{KL}\left(p_\phi(\mathbf{z}) \,\|\, p_\xi(\mathbf{z})\right) - \beta\mathbb{E}_{p(\mathbf{x})}\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \log p_\theta(\mathbf{x}\mid\mathbf{z})\end{aligned}$$

Finally they introduce an estimate of the true data density $p_\varphi(\mathbf{x}) \approx p(\mathbf{x})$ and simply plug into the objective a KL divergence to minimize the distance between the two (which is equivalent to maximizing the log likelihood $\mathbb{E}_{p(\mathbf{x})} \log p_\varphi(\mathbf{x})$) making the objective yet another upper bound on the original (we minimize the objective)

$$\begin{aligned}&\mathcal{L}^{U}[q_\phi(\mathbf{z}\mid\mathbf{x})] \le \mathcal{L}^{UB}[q_\phi(\mathbf{z}\mid\mathbf{x}), p_\theta(\mathbf{x}\mid\mathbf{z})] \le \mathcal{L}^{UBD}[q_\phi(\mathbf{z}\mid\mathbf{x}), p_\theta(\mathbf{x}\mid\mathbf{z}), p_\xi(\mathbf{z}), p_\varphi(\mathbf{x})] := \\ &\underbrace{\mathbb{E}_{p(\mathbf{x})}\mathrm{KL}\left(q_\phi(\mathbf{z}\mid\mathbf{x}) \,\|\, p_\xi(\mathbf{z})\right)}_{A} \underbrace{- \mathrm{KL}\left(p_\phi(\mathbf{z}) \,\|\, p_\xi(\mathbf{z})\right)}_{B} \underbrace{- \beta\,\mathbb{E}_{p(\mathbf{x})}\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \log p_\theta(\mathbf{x}\mid\mathbf{z})}_{C} \underbrace{+ \beta\,\mathrm{KL}\left(p(\mathbf{x}) \,\|\, p_\varphi(\mathbf{x})\right)}_{D}\end{aligned}$$

> This is a very weird step. The only link between the approximate data density $p_\varphi(\mathbf{x})$ and all the other learned distributions is only through the variational objective, not through basic probability manipulations as is usual in VAEs where we have $p_\theta(\mathbf{x}) = \int_{\mathbf{z}} p(\mathbf{z}) p_\theta(\mathbf{x}\mid\mathbf{z})\,\mathrm{d}\mathbf{z}$ which is estimated for each $\mathbf{x} \in \mathcal{X}$ by importance sampling $\int_{\mathbf{z}} q_\phi(\mathbf{z}\mid\mathbf{x}) \frac{p(\mathbf{z})}{q_\phi(\mathbf{z}|\mathbf{x})} p_\theta(\mathbf{x}\mid\mathbf{z})\,\mathrm{d}\mathbf{z} \approx \frac{1}{k}\sum_i^k \frac{p(\mathbf{z}_i)}{q_\phi(\mathbf{z}_i|\mathbf{x})} p_\theta(\mathbf{x}\mid\mathbf{z}_i)$. Also, for the variational inference to be tractable, we need to pick for the approximating distribution some simple class, e.g. a Gaussian. This would mean that the density of all possible observed datasets is modelled simply as a Gaussian (with pre-fixed dimensions) $p_\varphi(\mathbf{x})$.

Term C is the reconstruction loss term of the VAE ELBO, A is similar to the regularization shrinking to the prior (though here $p_\xi(\mathbf{z})$ is learned), B is similar to what was added in InfoVAE paper (Zhao et al. 2017) though again with prior instead of $p_\xi(\mathbf{z})$ and D is just a likelihood maximization term. They call this final formulation *bounded information bottleneck AE* (BIB-AE).

TLDR: They introduce even more variational approximations: one for the marginal of $\mathbf{Z}$ $p_\xi(\mathbf{z}) \approx p_\phi(\mathbf{z})$ and another for the data distribution $p_\varphi(\mathbf{x}) \approx p(\mathbf{x})$. Bringing the first into the objective makes it a function of another parameter (ξ), the data probability is brought simply by adding another KL term so that finally we have

$$\mathcal{L}^U[q_\phi(\mathbf{z}\mid\mathbf{x})] \le \mathcal{L}^{UB}[q_\phi(\mathbf{z}\mid\mathbf{x}), p_\theta(\mathbf{x}\mid\mathbf{z})] \le \mathcal{L}^{UBD}[q_\phi(\mathbf{z}\mid\mathbf{x}), p_\theta(\mathbf{x}\mid\mathbf{z}), p_\xi(\mathbf{z}), p_\varphi(\mathbf{x})] :=$$
$$\underbrace{\mathbb{E}_{p(\mathbf{x})}\operatorname{KL}\left(q_\phi(\mathbf{z}\mid\mathbf{x}) \parallel p_\xi(\mathbf{z})\right)}_{A} \underbrace{-\operatorname{KL}\left(p_\phi(\mathbf{z}) \parallel p_\xi(\mathbf{z})\right)}_{B} \underbrace{-\beta\,\mathbb{E}_{p(\mathbf{x})}\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\log p_\theta(\mathbf{x}\mid\mathbf{z})}_{C} \underbrace{+\beta\operatorname{KL}\left(p(\mathbf{x}) \parallel p_\varphi(\mathbf{x})\right)}_{D}$$

A and C are in elbo (except that here the prior $p_\xi(\mathbf{z})$ is learned), B is in InfoVAE and D is just maximizing the likelihood $p_\varphi(\mathbf{x})$. This they call the *bounded information bottleneck AE* (BIB-AE).

6.3.2 Links to generative adversarial model

They further try to link this to GANs but the treatment is so strange (e.g. assuming that we observe data-noise pairs $\{(\mathbf{x}_i, \mathbf{z}_i) \in \mathbb{R}^d \times \mathbb{R}^h\}_{i=1}^N$ and that we can actually compute directly some loss between the generator and the data such as $\mathbb{E}\|\mathbf{x} - g_\theta(\mathbf{z})\|$) that I won't discuss it further.

6.4 My rewriting the objective

I will drop β, replace the last two variational approximations $p_\xi(\mathbf{z})$ simply with a prior $p(\mathbf{z})$ and with the learned likelihood linked to the learned distributions through the standard VAE assumption $p_\varphi(\mathbf{x}) = p_\theta(\mathbf{x}) = \int_{\mathbf{z}} p(\mathbf{z})p_\theta(\mathbf{x}\mid\mathbf{z})\,\mathrm{d}\,\mathbf{z}$.

Standard results for ELBO in VAE say that

$$\begin{aligned}
ELBO &= \mathbb{E}_{p(\mathbf{x})}\log p_\theta(\mathbf{x}) - \mathbb{E}_{p(\mathbf{x})}\operatorname{KL}\left(q_\phi(\mathbf{z}\mid\mathbf{x}) \parallel p_\theta(\mathbf{z}\mid\mathbf{x})\right) \\
&= \mathbb{E}_{p(\mathbf{x})}\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\log p_\theta(\mathbf{x}\mid\mathbf{z}) - \mathbb{E}_{p(\mathbf{x})}\operatorname{KL}\left(q_\phi(\mathbf{z}\mid\mathbf{x}) \parallel p(\mathbf{z})\right) \\
&= \mathbb{E}_{p(\mathbf{x})}\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\log \frac{p_\theta(\mathbf{x},\mathbf{z})}{q_\phi(\mathbf{x},\mathbf{z})} + \mathbb{E}_{p(\mathbf{x})}\log p(\mathbf{x}) \\
&= \mathbb{E}_{p(\mathbf{x})}\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\log \frac{p_\theta(\mathbf{x}\mid\mathbf{z})p(\mathbf{z})}{q_\phi(\mathbf{x}\mid\mathbf{z})p_\phi(\mathbf{z})} + \mathbb{E}_{p(\mathbf{x})}\log p(\mathbf{x}) \\
&= -\mathbb{E}_{p_\phi(\mathbf{z})}\operatorname{KL}\left(q_\phi(\mathbf{x}\mid\mathbf{z}) \parallel p_\theta(\mathbf{x}\mid\mathbf{z})\right) - \operatorname{KL}\left(p_\phi(\mathbf{z}) \parallel p(\mathbf{z})\right) + \mathbb{E}_{p(\mathbf{x})}\log p(\mathbf{x}) \\
&= \mathbb{E}_{p(\mathbf{x})}\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\log \frac{p_\theta(\mathbf{z}\mid\mathbf{x})p_\theta(\mathbf{x})}{q_\phi(\mathbf{z}\mid\mathbf{x})p(\mathbf{x})} + \mathbb{E}_{p(\mathbf{x})}\log p(\mathbf{x}) \\
&= -\mathbb{E}_{p(\mathbf{x})}\operatorname{KL}\left(q_\phi(\mathbf{z}\mid\mathbf{x}) \parallel p_\theta(\mathbf{z}\mid\mathbf{x})\right) - \operatorname{KL}\left(p(\mathbf{x}) \parallel p_\theta(\mathbf{x})\right) + \mathbb{E}_{p(\mathbf{x})}\log p(\mathbf{x}) \ ,
\end{aligned}$$

where

$$p_\theta(\mathbf{z}\mid\mathbf{x}) = \frac{p(\mathbf{z})p_\theta(\mathbf{x}\mid\mathbf{z})}{p_\theta(\mathbf{x})} \qquad p_\theta(\mathbf{x},\mathbf{z}) = p_\theta(\mathbf{x}\mid\mathbf{z})p(\mathbf{z}) \qquad q_\phi(\mathbf{x},\mathbf{z}) = q_\phi(\mathbf{z}\mid\mathbf{x})p(\mathbf{x}) \tag{6.10}$$

The objective is

$$\begin{aligned}
&\mathcal{L}[q_\phi(\mathbf{z} \mid \mathbf{x}), p_\theta(\mathbf{x} \mid \mathbf{z})] = \\
&= -ELBO - \mathrm{KL}\left(p_\phi(\mathbf{z}) \parallel p(\mathbf{z})\right) + \mathbb{E}_{p(\mathbf{x})} \log p(\mathbf{x}) - \mathbb{E}_{p(\mathbf{x})} \log p_\theta(\mathbf{x}) \\
&= \mathbb{E}_{p_\phi(\mathbf{z})} \mathrm{KL}\left(q_\phi(\mathbf{x} \mid \mathbf{z}) \parallel p_\theta(\mathbf{x} \mid \mathbf{z})\right) + \mathrm{KL}\left(p_\phi(\mathbf{z}) \parallel p(\mathbf{z})\right) - \mathbb{E}_{p(\mathbf{x})} \log p(\mathbf{x}) \\
&\quad - \mathrm{KL}\left(p_\phi(\mathbf{z}) \parallel p(\mathbf{z})\right) + \mathbb{E}_{p(\mathbf{x})} \log p(\mathbf{x}) - \mathbb{E}_{p(\mathbf{x})} \log p_\theta(\mathbf{x}) \\
&\quad \mathbb{E}_{p_\phi(\mathbf{z})} \mathrm{KL}\left(q_\phi(\mathbf{x} \mid \mathbf{z}) \parallel p_\theta(\mathbf{x} \mid \mathbf{z})\right) - \mathbb{E}_{p(\mathbf{x})} \log p_\theta(\mathbf{x})
\end{aligned}$$

Result: Standard VAE maximizes the follwing lower bound the log likelihood (ELBO)

$$ELBO = \mathbb{E}_{p(\mathbf{x})} \log p_\theta(\mathbf{x}) - \mathbb{E}_{p(\mathbf{x})} \mathrm{KL}\left(q_\phi(\mathbf{z} \mid \mathbf{x}) \parallel p_\theta(\mathbf{z} \mid \mathbf{x})\right) \tag{6.11}$$

BIB-AE maximizes a different lower bound

$$\mathbb{E}_{p(\mathbf{x})} \log p_\theta(\mathbf{x}) - \mathbb{E}_{p_\phi(\mathbf{z})} \mathrm{KL}\left(q_\phi(\mathbf{x} \mid \mathbf{z}) \parallel p_\theta(\mathbf{x} \mid \mathbf{z})\right) \tag{6.12}$$

Can it be shown that this lower bound is somehow tighter? (I don't think so.) Also, ELBO has expectations with respect to $p(\mathbf{x})$ which can be approximated by MC, BIB-AE has expectation with respect to $q_\phi(\mathbf{z})$ which I don't know how to simplify to be able to evaluate and train.

Concluding remarks: The paper is not good - difficult to follow and decipher due to cluttered and unclear notation and not enough motivation for the individual steps. My interpretaiton and rewrites help it quite a bit. It suffers from usual problems of IB methods: deterministic vs stochastic netowrks and discrete vs continuos differential entropy. There are far too many variational approximations (and hence networks to learn) and there is no hint on how the training shall be performed. What is even a trainable form of the objecive? **However, it was a useful trigger to read on IB in DL.**

References

[1] Thomas M Cover and Joy A Thomas. *Elements of Information Theory*. en. Wiley, 2006 (cit. on p. 2).

[2] Ziv Goldfeld et al. "Estimating Information Flow in Deep Neural Networks". In: *ICML*. 2019 (cit. on p. 9).

[3] Andrew Michael Saxe et al. "On the Information Bottleneck Theory of Deep Learning". In: *ICLR*. 2018 (cit. on pp. 8, 9).

[4] Ravid Shwartz-Ziv and Naftali Tishby. "Opening the Black Box of Deep Neural Networks via Information". In: *arXiv:1703.00810 [cs]* (2017) (cit. on p. 9).

[5] Naftali Tishby. *Information Theory of Deep Learning - Naftali Tishby*. Yandex, 2017 (cit. on pp. 7, 8).

[6] Naftali Tishby, Fernando C. Pereira, and William Bialek. "The Information Bottleneck Method". In: *arXiv:physics/0004057* (2000) (cit. on pp. 2, 10).

[7] Naftali Tishby and Noga Zaslavsky. "Deep Learning and the Information Bottleneck Principle". In: *IEEE Information Theory Workshop (ITW)*. 2015 (cit. on pp. 5, 10).

[8] Slava Voloshynovskiy et al. "Information Bottleneck through Variational Glasses". In: *arXiv:1912.00830 [cs]* (2019) (cit. on p. 10).

[9] Shengjia Zhao, Jiaming Song, and Stefano Ermon. *InfoVAE: Information Maximizing Variational Autoencoders*. en. 2017 (cit. on p. 13).

Index